MVA ASSIGNMENT 9

MEMBER INFORMATION

Ruixin Yang (RUID: 197000459)

Aishwarya Senthilvel (RUID: 199001269)

CLASSIFICATION INTO TEST AND TRAINING DATA:

We split the bank dataset into 75% Training Data and 25% Test Data.

```
> # Lets cut the data into two parts
> smp_size_raw <- floor(0.75 * nrow(bank))
> train_ind_raw <- sample(nrow(bank), size = smp_size_raw)
> train_raw.df <- as.data.frame(bank[train_ind_raw, ])
> test_raw.df <- as.data.frame(bank[-train_ind_raw, ])
```

DISCRIMINATIVE ANALYSIS

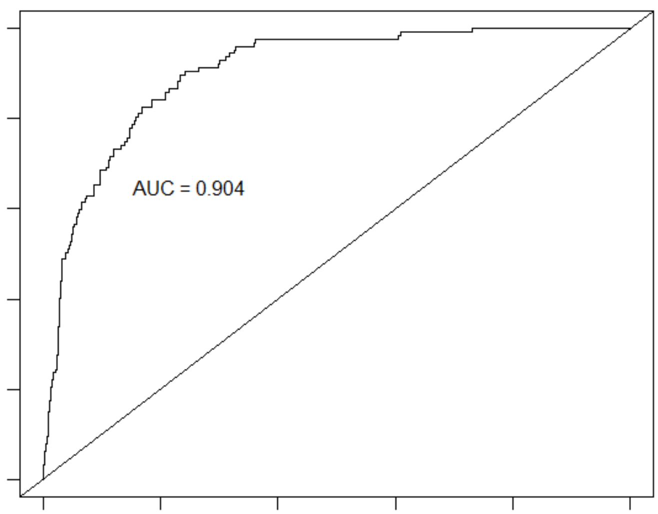
POSTERIOR PROBABLITY:

The first few lines of this output are shown below.

```
> # Get the posteriors as a dataframe.
> Bank_raw.lda.predict.posteriors <- as.data.frame(Bank_raw.lda.predict$posterior)
> Bank_raw.lda.predict.posteriors
               0            1
1    0.8427088326 0.1572911674
2    0.9895327139 0.0104672861
3    0.9612088229 0.0387911771
5    0.9964757053 0.0035242947
13   0.9543207973 0.0456792027
15   0.9983136299 0.0016863701
17   0.9743956061 0.0256043939
18   0.9662162548 0.0337837452
27   0.9959987727 0.0040012273
29   0.9942530240 0.0057469760
33   0.9945669655 0.0054330345
43   0.9713103363 0.0286896637
50   0.4074595779 0.5925404221
59   0.9817403415 0.0182596585
60   0.9487807492 0.0512192508
```

ROC/AUC CURVE:

```
> #create ROC/AUC curve
> library(ROCR)
> pred <- prediction(Bank_raw.lda.predict.posteriors[,2], test_raw.df$y)
> roc.perf = performance(pred, measure = "tpr", x.measure = "fpr")
> auc.train <- performance(pred, measure = "auc")
> auc.train <- auc.train@y.values
> plot(roc.perf)
> abline(a=0, b= 1)
> text(x = .25, y = .65 ,paste("AUC = ", round(auc.train[[1]],3), sep = ""))
```



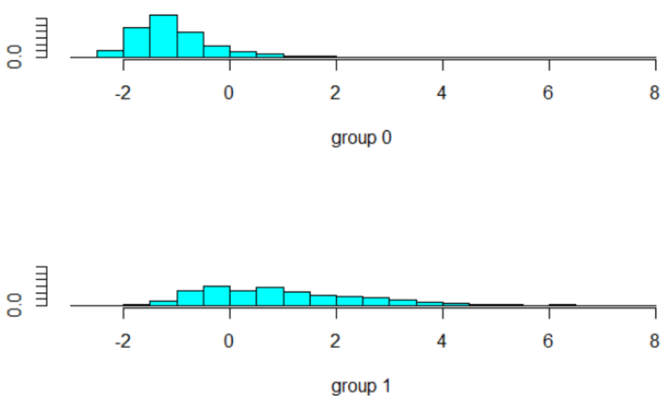
First, we ran a discriminative analysis to our dataset although our dependent variable is binary. And here is the summary of this analysis.

```
Call:
lda(y ~ ., data = bank)

Prior probabilities of groups:
      0       1 
0.88476 0.11524 

Group means:
  `Bank$age` `jobblue-collar` jobentrepreneur jobhousemaid jobmanagement jobretired `jobself-e
mployed`
0   40.99800       0.2192500      0.03825000    0.0245000     0.2095000  0.0440000         0.040750
00
1   42.49136       0.1324376      0.02879079    0.0268714     0.2514395  0.1036468         0.038387
72
  jobservices jobstudent jobtechnician jobunemployed jobunknown maritalmarried maritalsingle
```

```
0  0.09475000 0.01625000      0.171250    0.02875000  0.0077500      0.6300000     0.2572500
1  0.07293666 0.03646833      0.159309    0.02495202  0.0134357      0.5316699     0.3205374
  educationsecondary educationtertiary educationunknown    default `Bank$balance`   housing
  loan
0          0.5152500         0.2892500       0.04200000 0.01675000       1403.212 0.5847500 0.16200
000
1          0.4702495         0.3704415       0.03646833 0.01727447       1571.956 0.4222649 0.08253
359
  contacttelephone contactunknown `Bank$day`  monthaug   monthdec   monthfeb   monthjan  m
onthjul
0       0.06425000      0.3157500   15.94875 0.1385000 0.00275000 0.04600000 0.03300000 0.16
12500
1       0.08445298      0.1170825   15.65835 0.1516315 0.01727447 0.07293666 0.03071017 0.11
70825
   monthjun  monthmar  monthmay   monthnov   monthoct   monthsep duration campaign    pdays
 previous
0 0.1190000 0.0070000 0.3262500 0.08750000 0.01075000 0.00875000 226.3475 2.862250 36.0
0600 0.471250
1 0.1055662 0.0403071 0.1785029 0.07485605 0.07101727 0.03262956 552.7428 2.266795 68.6
3916 1.090211
  poutcomeother poutcomesuccess poutcomeunknown
0    0.03975000        0.011500        0.842000
1    0.07293666        0.159309        0.646833

Coefficients of linear discriminants:
                          LD1
`Bank$age`          1.019559e-03
`jobblue-collar`   -1.981906e-01
jobentrepreneur    -1.331150e-01
jobhousemaid       -2.136015e-01
jobmanagement      -5.651623e-02
jobretired          3.763192e-01
`jobself-employed` -7.170925e-02
jobservices        -8.490722e-02
jobstudent          3.389620e-01
jobtechnician      -1.293220e-01
jobunemployed      -3.307753e-01
jobunknown          2.891153e-01
maritalmarried     -2.719141e-01
maritalsingle      -1.527515e-01
educationsecondary  4.724354e-03
educationtertiary   1.379370e-01
educationunknown   -2.064974e-01
default             3.400915e-01
`Bank$balance`     -5.072348e-06
housing            -1.267259e-01
```

```
loan              -2.329421e-01
contacttelephone   3.362894e-02
contactunknown    -5.303392e-01
`Bank$day`         1.066742e-02
monthaug          -2.294423e-01
monthdec           3.379801e-01
monthfeb           1.136267e-01
monthjan          -7.127508e-01
monthjul          -4.463307e-01
monthjun           2.123845e-01
monthmar           1.518184e+00
monthmay          -2.356932e-01
monthnov          -4.890867e-01
monthoct           1.562553e+00
monthsep           6.724373e-01
duration           3.332826e-03
campaign          -9.865886e-03
pdays             -3.605995e-04
previous          -3.428844e-03
poutcomeother      3.719580e-01
poutcomesuccess    2.895548e+00
poutcomeunknown   -1.667129e-01
```

We also made a figure about this analysis. Here is the outcome.



Besides, we also get to know that numbers of 0 and 1 are 4000 and 521 in our dataset. And the ratio of between- and within- group standard deviations is 42.97. Its square is the canonical F statistics. Then, we used the leave-one-out cross-validation to our discriminative analysis. Based on it, we then calculated the posterior probabilities. Here is its descriptive statistical result.

0	1
Min. :0.0000005	Min. :0.0005829
1st Qu.:0.9479935	1st Qu.:0.0063456
Median :0.9855353	Median :0.0144647
Mean :0.8931941	Mean :0.1068059
3rd Qu.:0.9936544	3rd Qu.:0.0520065
Max. :0.9994171	Max. :0.9999995

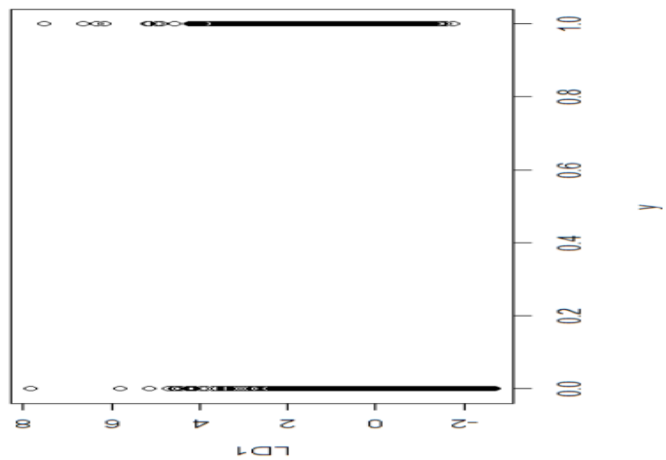
The leave-one-out cross validation uses only one observation as the test subset. We decided to do another cross validation with half of the sample as the test subset. Additionally, we also set that the prior probabilities of 0 and 1 as 0.5. Then we made predictions based on this model and compare the prediction with the test group. First, we got to know that numbers of 0 and 1 are 1886 and 375 in the prediction set, respectively. That comes from posterior probabilities. Here is the descriptive statistical report of posterior probabilities.

0	1
Min. :0.0000002	Min. :0.004152
1st Qu.:0.7314263	1st Qu.:0.045578
Median :0.9050697	Median :0.094930
Mean :0.7695441	Mean :0.230456
3rd Qu.:0.9544216	3rd Qu.:0.268574
Max. :0.9958478	Max. :1.000000

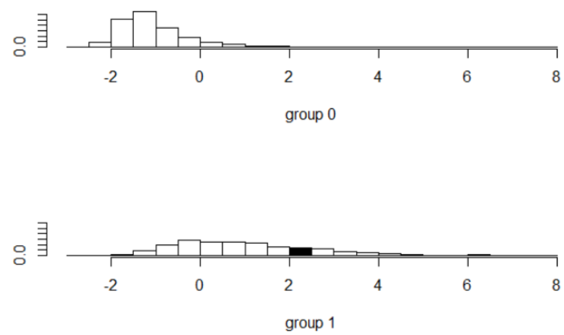
Then, we set the prior probabilities as both 0.5 and ran the analysis again without cross validation. The result is the same as that of the first analysis. From this analysis, we later made predictions about the whole dataset. And here is the descriptive statistical summary.

0	1
Min. :0.0000002	Min. :0.004646
1st Qu.:0.7053667	1st Qu.:0.046753
Median :0.8992097	Median :0.100790
Mean :0.7605423	Mean :0.239458
3rd Qu.:0.9532466	3rd Qu.:0.294633
Max. :0.9953541	Max. :1.000000

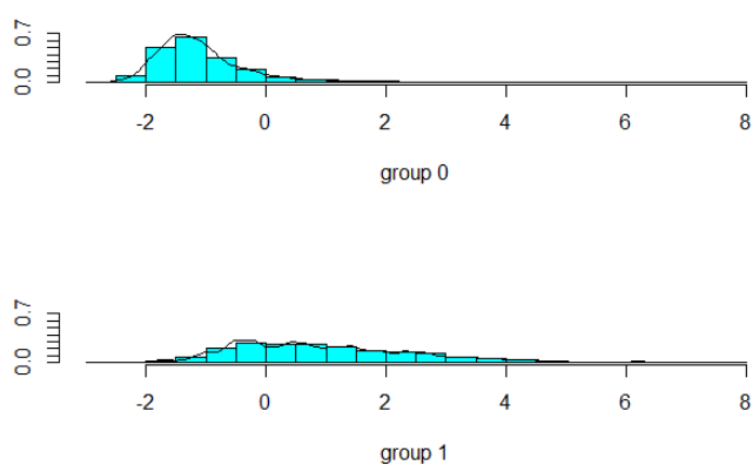
Since our dependent variable is binary, we made the figure containing distributions of observations in each group. Here is the result.



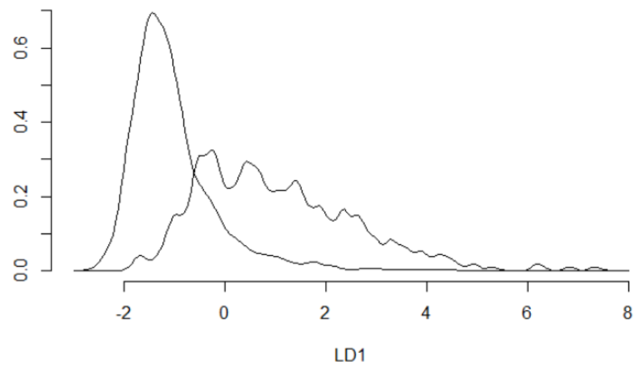
The upper dots are in group 1 while the lower dots are in group 0. This figure is actually correspondent to the figure with two histograms above since these two analyses have the same outcome. After this, we next drew a random sample as the training sample from the dataset. Its size is three quarters of the dataset. We again drew histograms of both groups. Here is the result.



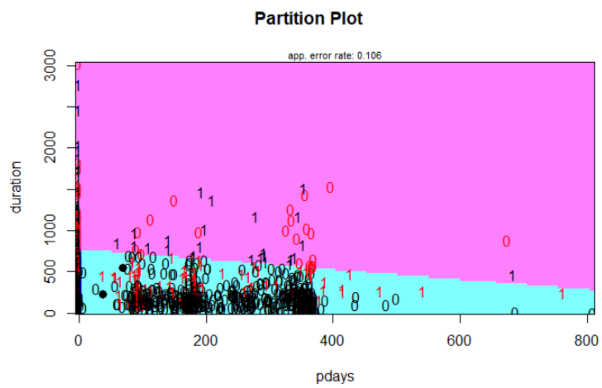
This one is just like the one with two histograms above. Next, we add smoothing curves and color the this bars. Here is the result.



Apart from these two figures above, we then made the density figure alone. Here it is.



The curve with higher summit and higher positive skewness represents the group of zeros. The other one represents the group of ones. Next, we got the partition plot for numeric variables within our dataset. But the size is too large. So, we only got the one for duration and pdays variables.



As you can see, these two groups overlap highly. Then, we made predictions from the outcome from the training set and compare it with the truth of the training set. Here is the table of accuracy. But this is an in-sample accuracy table.

	0	1
0	2887	212
1	110	170

It is time for us to use our test set. We want to know the out-of-sample accuracy from the test set. Here is the table.

	0	1
0	968	78
1	35	61

At last, we employed the Wilk's Lambda test and F test. We applied these tests to a model with four numeric regressors and our only dependent variable. First, we show the result of Wilk's lambda test.

```
> summary(m,test="Wilks")
          Df   Wilks approx F num Df den Df    Pr(>F)
y          1 0.82654   236.94      4   4516 < 2.2e-16 ***
```

```
Residuals 4519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As you can see, the p value is very small. We intend to reject the null hypothesis. So, we think the overall model is significant. Then we show the result of F test.

```
summary(m,test="Pillai")
           Df  Pillai approx F num Df den Df    Pr(>F)    
y           1 0.17346   236.94      4   4516 < 2.2e-16 ***
Residuals 4519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It shows nearly the same result. Both reject the null hypothesis. Both indicate the model is significant in general. Finally, we attached the analysis of variances for this model.

```
> summary.aov(m)
 Response 1 :
             Df Sum Sq Mean Sq F value   Pr(>F)   
y             1   1028 1028.00  9.2071 0.002425 **
Residuals   4519 504562  111.65
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
             Df     Sum Sq  Mean Sq F value Pr(>F)
y             1 1.3126e+07 13125638  1.4492 0.2287
Residuals   4519 4.0929e+10  9057022

 Response duration :
             Df    Sum Sq  Mean Sq F value    Pr(>F)    
y             1  49107860 49107860  866.51 < 2.2e-16 ***
Residuals   4519 256107263    56673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response pdays :
             Df   Sum Sq Mean Sq F value    Pr(>F)    
y             1   490887  490887  49.495 2.287e-12 ***
Residuals   4519 44818676    9918
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

These tests are for the four numeric variables. Only the second regressor is insignificant while others are relatively significant.